Pessimistic Uplift Modeling

Atef Shaar
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay
Paris, France
atef.shaar@telecom-paristech.fr

Talel Abdessalem
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay
Paris, France
talel.abdessalem@telecom-paristech.fr

Olivier Segard
Télécom Ecole de Management, Institut Mines-Télécom
Evry, France
olivier.segard@telecom-em.eu

ABSTRACT

Uplift modeling is a machine learning technique that aims to model treatment effects heterogeneity. It has been used in business and health sectors to predict the effect of a specific action on a given individual. Despite its advantages, uplift models show high sensitivity to noise and disturbance, which leads to unreliable results. In this paper we show different approaches to address the problem of uplift modeling, we demonstrate how disturbance in data can affect uplift measurement. We propose a new approach, we call it *Pessimistic Uplift Modeling*, that minimizes disturbance effects. We compared our approach with the existing uplift methods, on simulated and real datasets. The experiments show that our approach outperforms the existing approaches, especially in the case of high noise data environment.

CCS Concepts

•**Computing methodologies** → *Classification and regression trees;* •**Applied computing** → *Marketing;* Health informatics; •**Mathematics of computing** → Causal networks;

Keywords

Uplift Modeling, Treatment effects heterogeneity, Deferential relational learning, Database Marketing

1. INTRODUCTION

The ambition of gauging the real impact of specific action on human behavior has always been a matter of debate, whether you are in business or in medical sector, the knowledge of how an action influences a desired behavior is good. Of course the possibility to predict a complicated and complex system as human decision system with all the technological advancement we have today is very limited. In the middle of this limitation and complication, uplift modeling appears to be the most reliable modeling technique that is aimed to focus on how an action could alter human decision, and by finding those differences in an action effects, we could apply our method to predict future events.

Many approaches has been developed to address the problem of uplift, the core concept is to model the variability of a treatment impact over a controlled experiment population or what we call heterogeneous treatment effects. Although most of uplift modeling methods promise to perform well, but only few really do. With many names such as treatment effect heterogeneity, differential response, personalized treatment learning, net lift, true lift and uplift modeling. Most of them are subjective and customized to answer determined question based on research and market objectives. With the lack of a general measurement tool, as we cannot apply and not apply a treatment to an individual, only simulation data seems to give us a workaround to solve this problem. Although successful model in simulation data does not guarantee a success in a real data, but it is the most trustworthy testing measurement for uplift modeling. A success on simulation and real datasets will be a significant indicator of the model performance.

Modeling treatment heterogeneity suffers from high risk of over-fitting, noise modeling, variable correlation and disturbance, it is crucial to address those problem while building an uplift modeling method, Leo Guelman[3] has successfully created two personalized treatment learning approaches that proposed a solution for those problems. Guelman's approaches is created based on a modified predictive model, but still shows some passive sensibility to high noise data and correlated features.

this motivates us to research for a solution that could avoid noise sensibility, offers accurate predictive scores and could be applied using any convenient predictive algorithm. By looking at the problem from another perspective, we focused on building a new method, we call it *Reflective Uplift* that is less sensitive to the noise in data. we developed a new approach by using the *Reflective Uplift* as a stabilizer to our Model, in this case we catch more uplift effects inside noisy environment

In this paper we will cover the evolution of modeling treatment heterogeneous effects, different types and approaches of uplift modeling, we will discuss its weak points and the attempts that has been made to address those weaknesses. Then we will introduce our new approach, we will demonstrate its advantages over other approaches, and how it will provide a better and more reliable uplift predictions. We will show how our modeling technique outperforms others, specially in noisy data-set using experiments on simulation

ACM SIGKDD '16 August-2016 San Francisco, California USA

DOI: 10.475/123_4

and real word datasets.

2. RELATED WORK

Uplift modeling tries to predict the effect of specific action on personal level. Usually two random samples required for uplift modeling, One sample called *Treatment Sample*, they receive an action, another monitored sample called *Control Sample*, they don't receive the action, then an uplift score would be measured by calculating the difference between predicted probabilities of each samples' predictive model. Most of uplift literature used this method in marketing domain, specifically it has been used to increase the effectiveness of marketing campaigns. An intuitive way for uplift modeling would follow its definition by building two predictive response models, one model would be built based on treatment sample, the second using control sample, then we get predicted uplift by simply subtract the two predicted probabilities.

$$Uplift = P(R|T, x) - P(R|C, x)$$

Uplift models goal is to classify individuals into four quarters(see Figure 1).

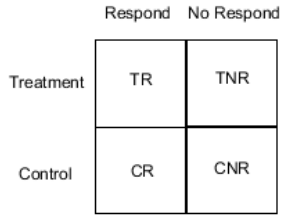

Figure 1: contingency table of uplift modeling

We can categorize uplift modeling techniques into direct and indirect methods. Direct uplift methods build only one model using the whole population to predict uplift. While indirect uplift methods build two separated models, one for each sample, then calculates uplift score as we mentioned earlier(see Figure2). The first uplift model by Radcliffe et

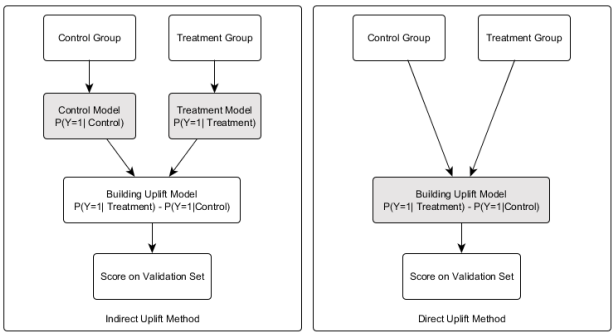

Figure 2: Direct Vs. Indirect Uplift Modeling

al.[15] was a direct model using a modified *CART* (Classification and Regression Trees)Breiman 1984[2]. Various predictive methods have been used to build uplift models, Logisitc regression [4, 10], Neural Network [12], Decision Trees[4, 15, 17, 3], KNN[1, 20] and SVM [6]. Almost all direct uplift methods need to develop a modified version of a predictive model and customize it to predict uplift effects. On the other hand, all indirect models used convenient predictive methods to build uplift models. There is a trade-off between the simplicity in indirect methods and its accuracy. Despite that trade-off, there is one special method proposed by Lai[9], this method has higher predictability score comparing to other indirect method. *Lai's Uplift Method* was based on the idea of Positive and Negative Uplift. Lai argued that positive uplift will be inside the diagonal of *Treated Respond* and *Controlled Not Respond*, the gray areas in Figure 3. While the negative uplift is in the white area, in *Treated Not Respond* and *Controlled Respond*. After that a classifier has to be built to predict the probability of a new case to be in positive uplift.

$$\textit{Lai's method Uplift} = P(TR + CNR) - P(TNR + CR)$$

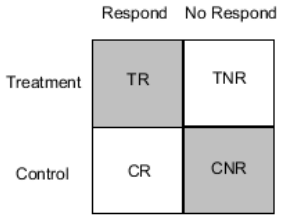

Figure 3: Lai's approach contingency table of uplift modeling; gray areas represent *positive uplift*; white areas represent *negative uplift.*

In 2014 Kane et. al[8] proofed mathematically that *Lai's uplift method* is a special case of more general equation that has been introduced in Kane's paper, which is Lai's generalized method.

$$\textit{Lai's generalized Uplift} = \frac{P(TR) - P(TNR)}{P(T)} + \frac{P(CNR) - P(CR)}{P(C)}$$

Uplift models is a classification model, it tends to be very sensitive to noise in data. Indirect methods are more affected by noise[16] than direct methods. Despite its importance for uplift modeling, few have researched how noisy data could affect uplift model performance. In general there are two types of noise in data, *Class Noise* and *Attribute Noise*[23]. Class Noise happens because of contradictory and misclassified cases. Contradictory cases, when two different cases has same attributes values, but opposite classes, it is common noise with uplift data-sets, basically because consumer behavior causality chain is much more complicated to be naively characterized by a set of unique features. Misclassified cases, also common with uplift, because in general missing an action is recorded as not an action, It could happen because of monitoring error, sometimes because users have some sort of a technique to prevent monitoring the desired action, in this case even if the individual did an action, a "no action" will be recorded instead. Attribute noise which is more harmful for a classifier performance than class noise[23], happens when there is missing or Incorrect attribute data, it is a major problem for any classifier

that deals with personal data, it is related to the individual willingness to share data, this has a huge impact on uplift modeling, sometimes the effect of an action will lead to not sharing the data[11]. An Uplift model performance will be affected also by the correlation between attributes and by the fact that uplift effect of a treatment could be so much weaker than the real treatment effect[3, 21]. Leo Guelman[3] addressed those issues in uplift modeling and created two new uplift models *Uplift Random Forests* and *Causal Conditional Inference Forests*. Guelman showed that *CCIF* method outperformed other uplift methods.

3. CONTRIBUTION

We introduce a new method for uplift modeling, *reflective uplift* , it provides an uplift score that is much less affected by *Response Disturbances*, it performs very well in supporting convenient uplift modeling technique to create a stable, less sensitive and reliable uplift model.

As we mentioned before uplift modeling sensitivity to noise could minimize its importance and reliably. A good uplift method would consider solving those issues. We want to solve uplift model problems using convenient predictive models. Looking as the four classification areas of uplift, we can see that there is two disturbance effect that could increase misclassification error. Main treatment effect or *Response effect* will be between two groups *Responders* and *Non Responders*(see Figure4), this effect happened because people who response usually have similarities, this effect is what a *Response model* will model. Usually it is much more stronger than the uplift effect. The second disturbance effect will be the *Partitioning Effect*, this will be between *treatment* and *control* areas, theoretically there should be no Partitioning Effect, but it is common to find this effect, mainly because human biased intervention. This effect could be minimized by proper sampling and by applying bagging algorithms.

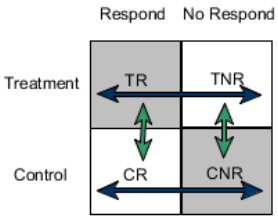

Figure 4: Uplift Disturbance Effects
Response effect: horizontal; Partitioning effect: vertical
Randomized and proper sampling will minimize partitioning effect.

A simple example of response disturbance is shown in the table 1, the table include imaginary data for ten clients of email marketing campaign. We have four columns, a binary treatment that indicate whether or not the client receive an email, a binary response and two predictors. If we want building an uplift decision tree and we want to split based (*Coupon*2) variable, this variable represents whether the client used coupon or not on his last purchase, obviously it will correlate even a little with the response variable. When we split based on *Coupon*, we will get split

Treat	Response	Gender	Coupon
1	1	male	Yes
0	1	female	Yes
1	1	male	Yes
1	0	female	No
0	0	male	No
0	1	female	Yes
1	1	male	Yes
1	1	male	Yes
0	0	female	No
0	0	female	No

Table 1: Email Marketing Campaign Data. Note that Coupon variable is correlated with Response variable, which will lead to response disturbance.

based on *Response*, leading to wrong or disturbed predictions(see Figure 5). This example is an extreme case, usually there is no such a highly correlated variable, there is a set of correlated variables. The fact that responders usually share the same characteristics entails there will be variables that correlate with the response. A response model will use them for scoring, while an uplift model will try to find differences between those variables to differentiate between treatment and control responders. This is why we cannot fix respond disturbance by random sampling as we do with partitioning disturbance, respond disturbance contains an uplift informative values. In general response disturbance variables in data will lead to bias in splitting criteria and to an incorrect prediction score.

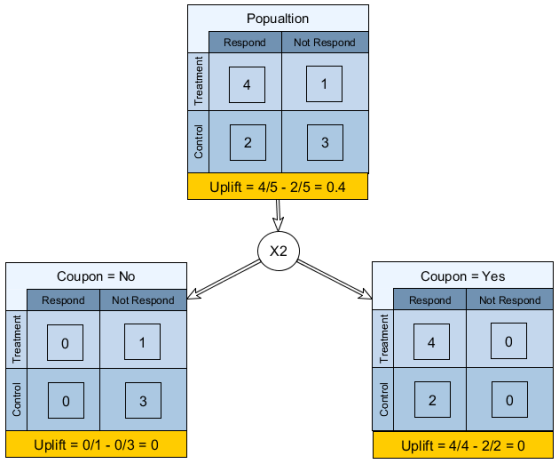

Figure 5: Uplift Decision Tree Node Split
Split based on respond correlated variable (Coupon).Uplift decision trees that use Uplift as stopping criteria will be the most affected.

We favored *Lai's uplift method* among others as it is less biased toward disturbances, mainly because of the diagonal partitioning of the population. Lai's method is simple and

effective, it can be applied easily using any binary classifier method. Also we found it logical to combine the negative uplift together, because individuals in those areas (*TNR* and *CR*) already made their mind *Not to* do the desired action under their circumstances. This made lai's method is a very good candidate to predict out basic uplift. We weighted lai's method by multiplying predicted probabilities by its cases proportions.

$$Uplift^{lai} = P(Positive|x)*P(\frac{positive}{population})-P(Negative|x)*P(\frac{negative}{population})$$

Algorithm 1 Lai's Uplift Model

1: Partition the population into two classes:
2: Positive Class: TR or CNR => Class = 1
3: Negative Class: TNR or CR => Class = 0
4: Build binary classification model with Class as a target variable
5: Positive Lift= $P(Class_{positive}|x_i) * \frac{positive\ cases}{population}$
6: Negative Lift= $P(Class_{negative}|x_i) * \frac{negative\ cases}{population}$
7: $Uplift_i^{lai} = Positife\ Lift - Negative\ Lift$
8: where x_i is a vector of features for individual i

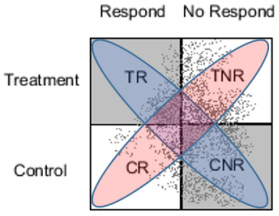

Figure 6: Lai's Uplift Partitions
TR and CNR is Positive; TNR and CR is NegativeNote that lai's method distribute the disturbance effects on the two partitions, so it minimize some of the disturbance

by applying on our example data set, we get $P(positive|Coupon = Yes) = 0.366$, while the uplift score by its definition $\frac{P(TR|coupon=Yes)}{P(T)} - \frac{P(CR|coupon=Yes)}{P(C)} = \frac{4}{4} - \frac{2}{2} = 0$.

We need an uplift model that should not affected by response disturbance biased, we need a model that can be used as a stabilizer to reduce disturbance sensitivity. This motivated us to build the *reflective uplift model*, it is a modified two model approach uplift modeling, it does not predict the probability of a case to respond after treatment, actually *reflective uplift* answer the question of what is the probability for a specific case to be in treatment area, given it responded. In this way we minimize the effect of respond disturbance. On the other hand we are aware that we maximized the sensitivity to partitioning disturbance, but as we mentioned earlier, portioning sensitivity is biased that can be avoided by proper sampling.

To build our *reflective uplift* model. First we partition the dataset into Respond and Not Respond, then we train a first model to predict the probability *Treated Respond* in *Responders* group, then we train the second model to predict the probability of *Treated Not Respond* in *Non Responders* group. Remember that $P(CR|R) = 1 - P(TR|R)$ and $P(CNR|R) = 1 - P(TNR|R)$. SO we calculate uplift by subtracting the negative uplift from the positive uplift.

$$P^{Reflective}(Positive|x) = P_{model}(T|R) * P(TR) + P_{model}(CNR|NR) * P(CNR)$$

$$P^{Reflective}(Negative|x) = P_{model}(T|NR) * P(TNR) + P_{model}(C|R) * P(CR)$$

$$Uplift^{reflective} = P^{Reflective}(Positive|x) - P^{Reflective}(Negative|x)$$

Algorithm 2 Reflective Uplift

1: Partition the Responders into two classes:
2: Treatment Class: TR => Class = 1
3: Control Class: CR => Class = 0
4: Build binary classification model with Class as a target variable
5: $M_i^{tr} = P(Class_{treatment}|R, x_i)$
6: Partition the Non Responders into two classes:
7: Treatment Class: TNR => Class = 1
8: Control Class: CNR => Class = 0
9: Build binary classification model with Class as a target variable
10: $M_i^{tnr} = P(Class_{treatment}|NR, x_i)$
11: Positive Uplift = $M_i^{tr} * P(TR) + (1 - M_i^{tnr}) * P(CNR)$
12: Negative Uplift = $M_i^{tnr} * P(TNR) + (1 - M_i^{tr}) * P(CR)$
13: $Uplift_i^{Reflective} = Positive\ Uplift - Negative\ Uplift$
14: where x_i is a vector of features for individual i

If we apply *reflective uplift* to predict $P(positive|Coupon = Yes)$, we get $Uplift^{reflective} = 0.266$, which makes sense, as the number of treated respondents are bigger than the control once. We applied ensemble decision trees methods while building each model to minimized the over-fitting and misclassification error rate, ensemble methods showed great improvement in uplift model performance as Soltys et al. 2014[19] and Leo Guelman[3] demonstrated. We Calculated our final uplift score as followed

$$Pessimistic\ Uplift = 1/2 * (Uplift_i^{lai} + Uplit_i^{Reflective})$$

By combining *Lai's* with *Reflected*, we got *Pessimistic Uplift*, an equilibrium predictive model that have precision, robustness and reliability.

4. EXPERIMENTS

We followed the steps of Leo Guelman[3] experiments,which is a modified version of Tian et. al experiment framework[21]. We evaluated our method within eight different scenarios, scenarios differs by the factor of strength of the main treatment effect over heterogeneity effect, the correlations between features and the magnitude of noise in the simulated data. Using simulation data for evaluation gave us the advantage of comparing uplift model score with the real personal uplift, this advantage that cannot be found in real data.

We benchmark models performance using spearman rank correlation between the real and predicted uplift. We used

Uplift R package, developed by Guelman[3] to produce the simulated data, also we used the same package for applying the CCIF (*Causal Conditional Inference Forests*) method, we compared our method to *Two Models method* ($Uplift = P(TR/T) - P(CR/C)$)[10], generalized *Lai's method*, as proposed by Kane et. al[8] and CCIF[3] with its default settings.We used ensemble methods to build all the models but CCIF. It is important to compare with CCIF as CCIF is the most concrete method based on Guelman[3].

The first four scenarios has the main effect twice as big as the uplift effect (treatment heterogeneity), while the last four scenarios has the main effect four times bigger than the uplift effect. The correlation among features variates between 0 and 0.5, also the magnitude of noise switch between $\sqrt{2}$ and $2\sqrt{2}$ respectively. For more details about the simulation framework, you can check Guelman (2014)[3] and Tian (2014)[21].

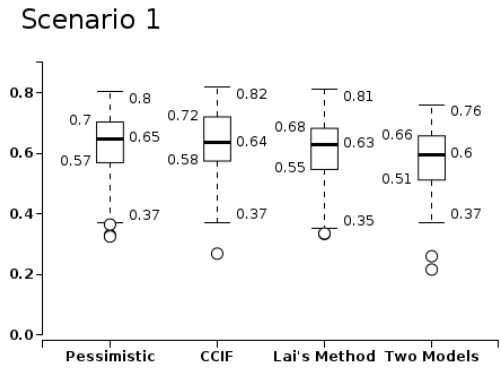

Figure 7: Scenarios 1: Main effect strength: Low; Correlation among features: Low; Magnitude of noise: Low

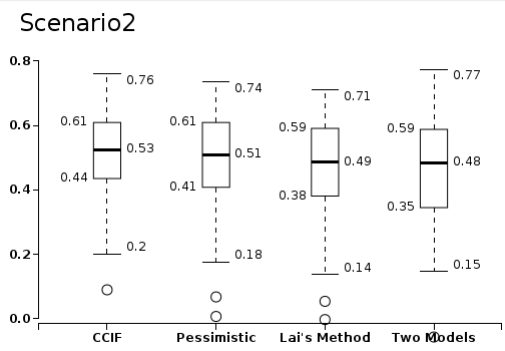

Figure 8: Scenarios 2: Main effect strength: Low; Correlation among features: Low; Magnitude of noise: High

The training data set contains 200 rows, and 10000 rows for validation dataset. Each dataset contains treatment column and response column, both of them are binary variables. Datasets contains twenty features named X1 to X20, only X1, X2, X3 and X4 has treatment heterogeneity effect. We build our models using *Ensemble Regression Trees*, one hundred ensemble trees, we included all features in building our models, with a fraction of 80% of training data for learning in each tree model and minimum 20 rows in each node split. We repeat each experiment one hundred times

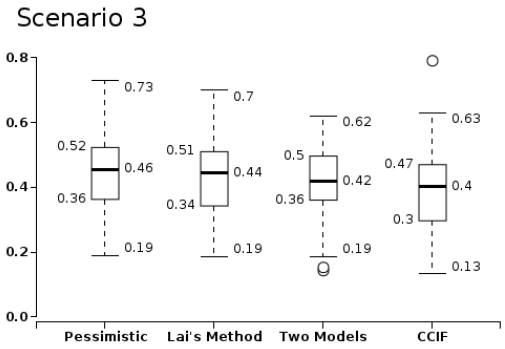

Figure 9: Scenarios 3: Main effect strength: Low; Correlation among features: High; Magnitude of noise: Low

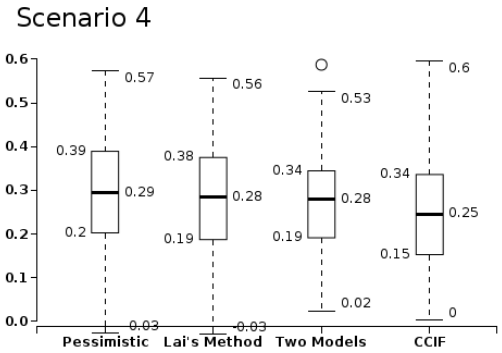

Figure 10: Scenarios 4: Main effect strength: Low; Correlation among features: High; Magnitude of noise: High

for each scenario. We measure model performance using Spearman's rank correlation coefficient between the real uplift and the predicted uplift. As shown in the Figure 7 , each box-plot represent the average Spearman's rank correlation coefficient for that model.

We can see how our model almost scores the same as CCIF in the first, second, fifth and sixth,but our model outperforms CCIf in the third, fourth, seventh and eighth scenarios, those scenarios has more correlations between features, with the highest noise , our method performs better.

We did experiments on real data sets, the main problem of uplift models is the lack of reliable measurement tool, many measurements have been used, some used *Gini* and Top 15% Gini [8], others used an extension of *Gain Curve*, which is *Qini Curve* and *Qini coefficien* proposed by Nicholas Radcliffe[14], also Area Under Uplift Curve(AUUC) which is used by [7, 19, 17, 18, 22], some measurement techniques was subjected to the purpose of the uplift model, for example, authors in [16, 9] used net campaign profit to measure the success of am uplift model.

For the purpose of uplift model evaluation, we will compare models based on *Real Uplift Curve*(RUC), *precision* , and *effectiveness*. *Real Uplift Curve* shows the *Real uplift* of a population after applying the model on it. To draw RUC curve, we sorted our cases based on the model*Predicted Uplift* scores, then we binned that data into ten bins with equal frequency, so each bin holds 10% of the data. Then we calculated each bin's *Mean Predicted Uplift* and *Real uplift*, after

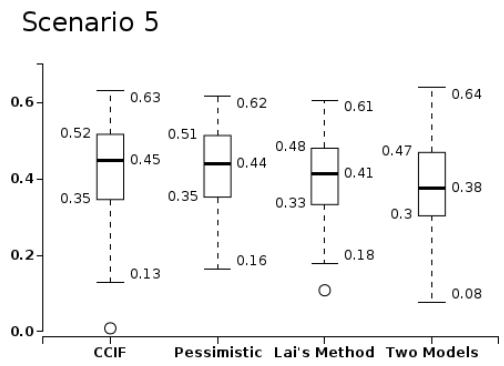

Figure 11: Scenarios 5: Main effect strength: High; Correlation among features: Low; Magnitude of noise: Low

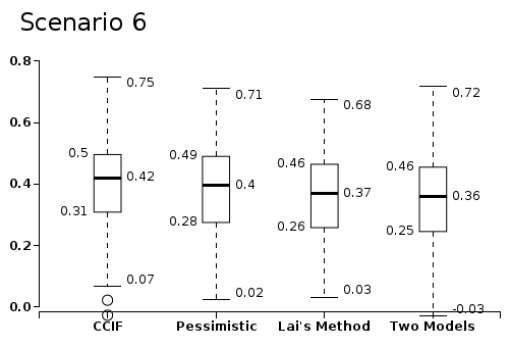

Figure 12: Scenarios 6: Main effect strength: High; Correlation among features: Low; Magnitude of noise: High

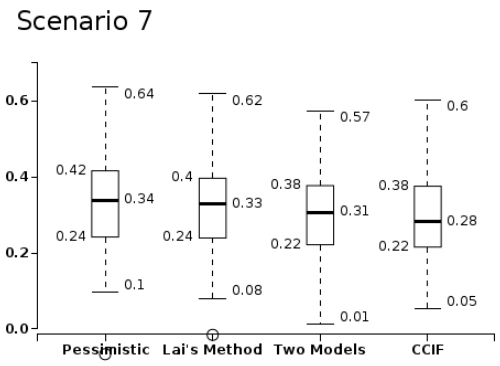

Figure 13: Scenarios 7: Main effect strength: High; Correlation among features: High; Magnitude of noise: Low

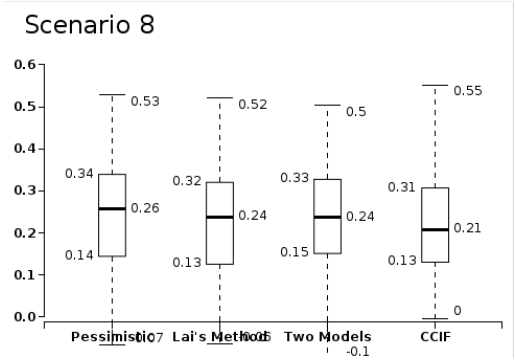

Figure 14: Scenarios 8: Main effect strength: High; Correlation among features: High; Magnitude of noise: High

that we draw the *Real Uplift Curve* where X axis represents the population and Y axis represents the *Real uplift* that we can get for each bin using the model.(see Figure15)

Model *precision* is measured by doing Spearman's rank correlation coefficient between *Mean Predicted Uplift* and *Real Uplift* (See Table3). Finally we calculate model *Effectiveness* by summing all real uplift values of bins above the *Random Uplift Line*(See Table2). *Random Uplift Line* is a straight line(see Figure15) represents the general uplift score that you will get by randomly sorting the population, it is calculated by uplift definition $P(TR|T) - P(CR|C)$. We used *Splice* and *Breast-Cancer* datasets from UCI Machine learning repository. We convert them to binary classification dataset. We followed the same way of binary conversion mentioned in [17, 18, 22, 19]. For *Splice* dataset, we chose treatment group when (*attribute_1*) attribute is equal to "A" or "G", while control group is when (*attribute_1*) attribute is equal to "T","D" or "C". A positive respond is when *Class* attribute equal "IE", and negative respond is when *Class* attribute equal "N".

For *Breast-Cancer* dataset, we used (*menopause*) attribute to partition into treatment and control group, and (*Class*) attribute as a target variable to predict.

We used *Hillstrom visit* dataset[5], this dataset contains data for an email marketing campaign experiment, it contains 64,000 different customers, spitted into three parts based on the type of email they got, Control Group got no-email, Women's Group and Men's Group. We used the dataset twice, once comparing Men's Group with Control, and another comparing Women's Group with Control.

We used Bone Marrow Transplant(BMT) and *Tamoxifen* datasets[13], BMT dataset is about patients who got transplant of bone marrow, there is two sources of bone marrow, pelvic bone or peripheral blood. We partition the dataset to treatment and control samples based on the source of bone marrow (pelvic bone as control). BMT has been used twice for experiments, once to predict the occurrence of acute graft versus host disease (agvh), second time to predict the occurrence of chronic graft versus host disease (cgvh). *Tamoxifen* dataset[13] contains experiment data for treating breast cancer using *Tamoxifen*, trying to predict if the person is alive or not, we split the dataset into control group, those who received *Tamoxifen* alone,and treatment group, who received Tamoxifen and Radio Therapy together.

From Our experiments on real datasets, we can see that both methods performed well by predicting uplift effects, *Pessimistic Approach* was more effective than *CCIF*, while *CCIF* shows some more *Precision* than *Pessimistic* specially with Hillstrom visit w. dataset.

From Our experiments on real datasets, we can see that both methods performed well by predicting uplift effects, *Pessimistic Approach* was more effective than *CCIF*, while *CCIF* shows some more *Precision* than *Pessimistic* specially with Hillstrom visit w. dataset.

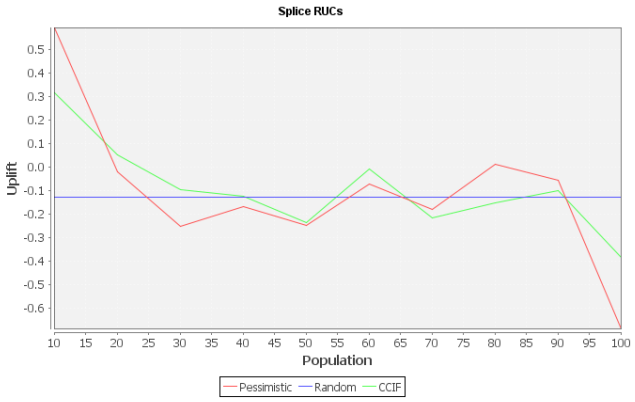

Figure 15: spliceRUC

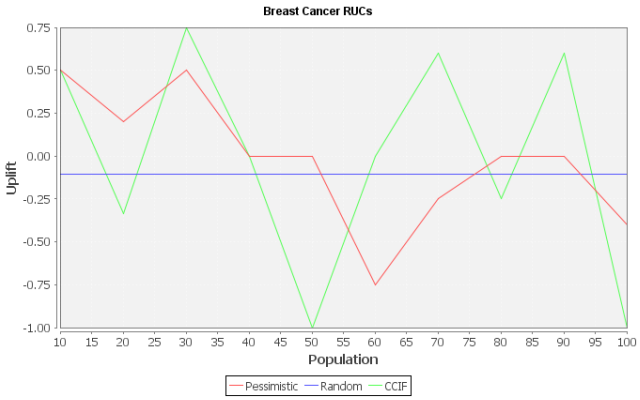

Figure 16: Breast-Cancer RUC

Dataset Name	*Random*	*CCIF*	*Pessimistic*
Splice(Artificial)	-0.12	0.08	**0.11**
Breast Cancer(Artificial)	-0.10	**0.28**	0.15
Tamoxifen(Real)	-0.01	0.02	**0.06**
Hillstrom visit(Real)	0.06	0.11	**0.12**
Hillstrom visit m.(Real)	0.06	**0.13**	**0.13**
Hillstrom visit w.(Real)	-0.05	**0.01**	**0.01**
BMT cgvh(Real)	-0.09	0.06	**0.10**
BMT agvh(Real)	-0.12	0.04	**0.11**

Table 2: CCIF vs Pessimistic Effectiveness
Effectiveness is calculated by summing all positive uplift above the random line;Each dataset was spitted into 80% training and 20% validation; Only BMT dataset has been used totally for training and validation because it contains only 100 rows

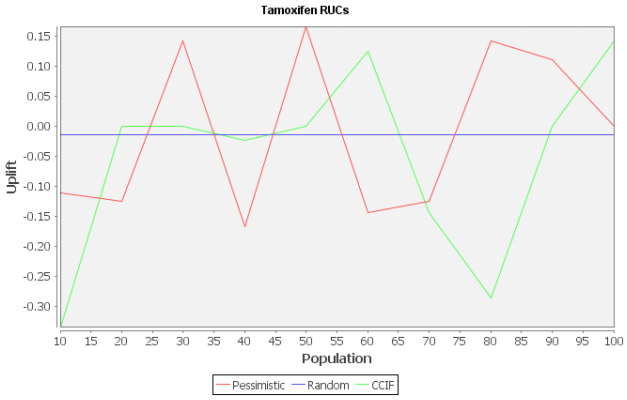

Figure 17: Tamoxifen RUC

Dataset Name	*CCIF*	*Pessimistic*
Splice(Artificial)	**0.70**	0.32
Breast Cancer(Artificial)	0.17	**0.73**
Tamoxifen(Real)	-0.36	**-0.20**
Hillstrom visit(Real)	**0.52**	0.34
Hillstrom visit m.(Real)	**-0.07**	-0.28
Hillstrom visit w.(Real)	**0.81**	0.00
BMT cgvh(Real)	**0.67**	0.52
BMT agvh(Real)	0.56	**0.68**

Table 3: CCIF vs Pessimistic Precision; Precision is calculated by doing spearman rank correlation between Mean Predicted Uplift and Real Uplift

Figure 18: Hillstrom Visit RUC

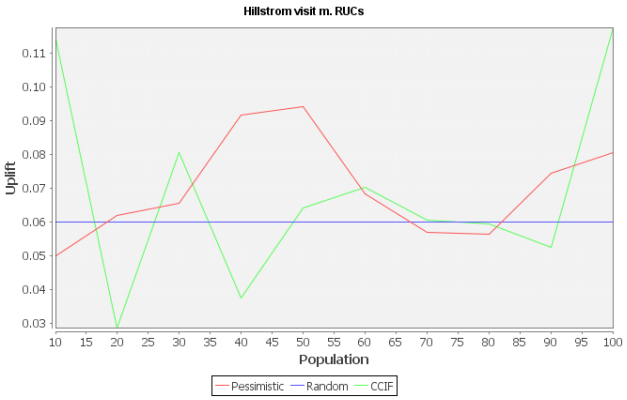

Figure 19: Hillstrom Visit m. RUC

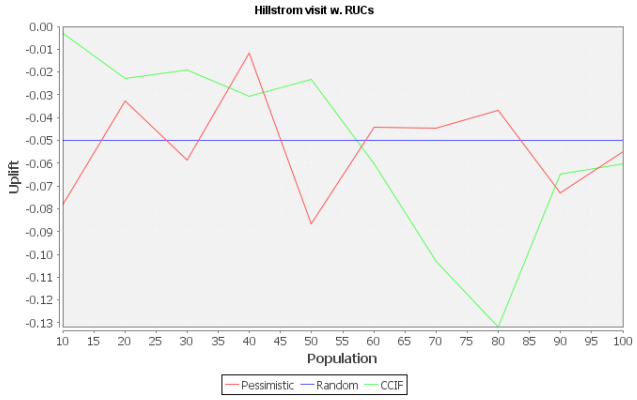

Figure 20: Hillstrom Visit w. RUC

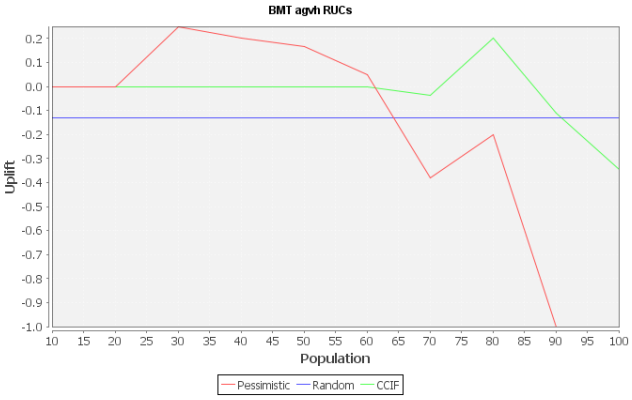

Figure 22: BMT agvh RUC

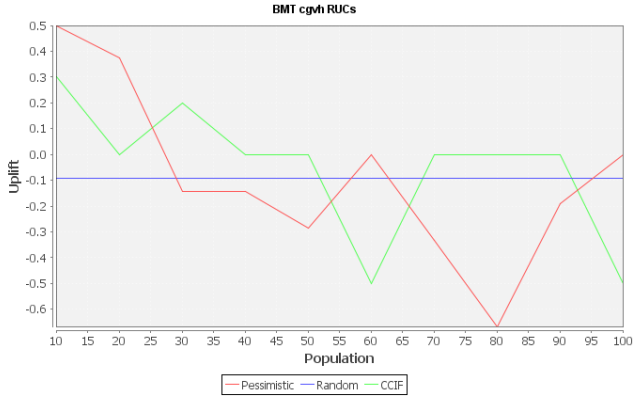

Figure 21: BMT cgvh RUC

5. CONCLUSION

Uplift modeling is a technique to predict the influence of a specific action on an individual behavior. There are many approaches to build uplift models, most of them struggle to produce reliable scores in noisy datasets. Most of real world datasets contains noise and disturbances, specially for uplift modeling, as uplift effects tend to be smaller than the real treatment effect. This leads us to introduce our approach which is based on building a stabilizer model that would helps in avoiding results that are based on noise. We compared our approach with other uplift methods, using simulated and real datasets. We used *Real Uplift Curves*, effectiveness and precession to measure models. Our model shows more stable and effective model than other approaches. More research needed, to extend our method multi-treatment multi-target uplift modeling.

6. REFERENCES

[1] F. Alemi, H. Erdman, I. Griva, and C. H. Evans. Improved statistical methods are needed to advance personalized medicine. *The open translational medicine journal*, 1:16, 2009.

[2] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[3] L. Guelman. *Optimal personalized treatment learning models with insurance applications*. PhD thesis, Universitat de Barcelona, 3 2015.

[4] B. Hansotia and B. Rukstales. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35–46, 2002.

[5] K. Hillstrome. The MineThatData E-Mail Analytics And Data Mining Challenge. http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html, 2008. [Online; accessed 10-Novemebr-2015].

[6] K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.*, 7(1):443–470, 03 2013.

[7] M. Jaskowski and S. Jaroszewicz. Uplift modeling for clinical trial data. *ICML Workshop on Clinical Data Analysis*, 2012.

[8] K. Kane, V. S. Lo, and J. Zheng. True-lift modeling: Comparison of methods. *J Market Anal*, 2(4):218–238, Dec 2014.

[9] L. Y.-T. Lai. *Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers*. PhD thesis, Citeseer, 2006.

[10] V. S. Lo. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86, 2002.

[11] L. Mamlouk and O. Segard. Big data and intrusiveness: Marketing issues. *Indian Journal of Science and Technology*, 8(S4):189–193, 2015.

[12] C. Manahan. A proportional hazards approach to campaign list selection. In *Proceedings of the thirtieth annual SAS users group international conference (SUGI), Philadelphia, PA*. SAS Institute Inc., 2005.

[13] M. Pintilie. *Competing risks: a practical perspective*, volume 58. John Wiley & Sons, 2006.

[14] N. Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, pages 14–21, 2007.

[15] N. Radcliffe and P. Surry. Differential response analysis: Modeling true response by isolating the effect of a single action. *Credit Scoring and Credit Control VI. Edinburgh, Scotland*, 1999.

[16] N. J. Radcliffe and P. D. Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, 2011.

[17] P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 441–450. IEEE, 2010.

[18] P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

[19] M. SoÅĆtys, S. Jaroszewicz, and P. Rzepakowski. Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, pages 1–29, 2014.

[20] X. Su, J. Kang, J. Fan, R. A. Levine, and X. Yan. Facilitating score and causal inference trees for large observational studies. *The Journal of Machine Learning Research*, 13(1):2955–2994, 2012.

[21] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

[22] L. Zaniewicz and S. Jaroszewicz. Support vector machines for uplift modeling. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, pages 131–138, Dec 2013.

[23] X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210.